Exploring Health Insurance Costs & Factors

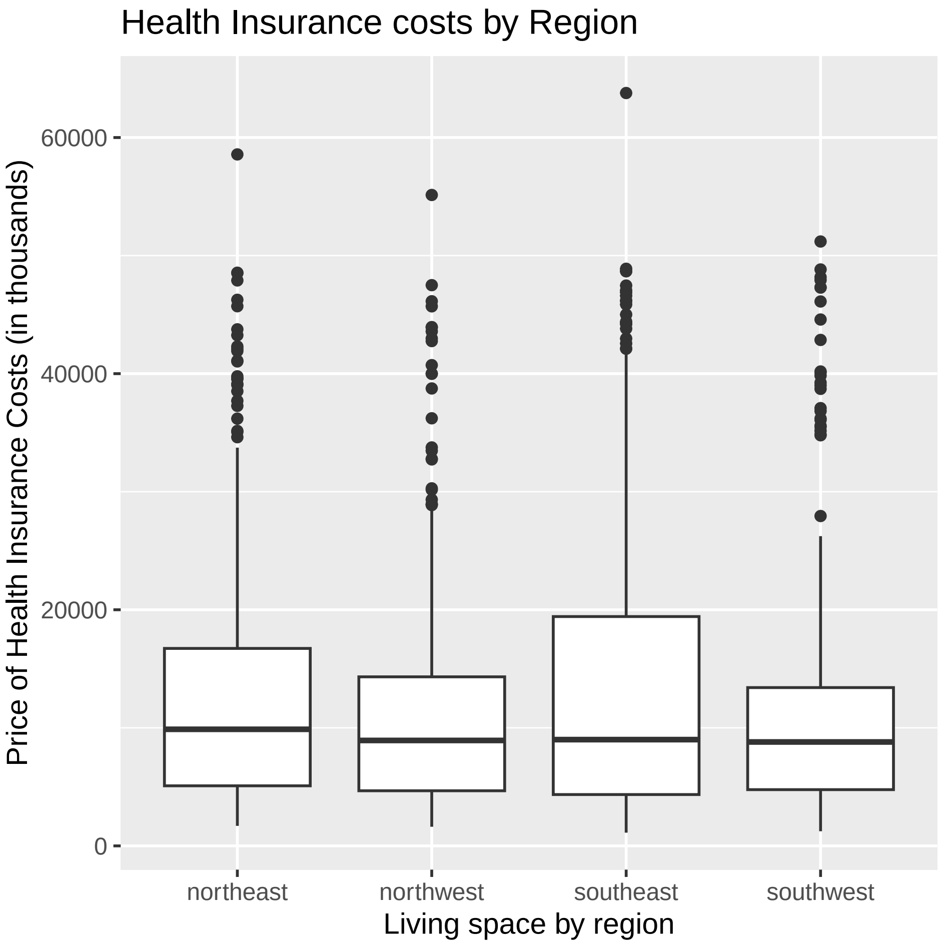
Acree Tejada

Introduction

In this study, I will be going over potential factors that influence cost of health insurance. Specifically, exploring two potential factors, region location and the number of children an individual has as predictors of their medical expenses. Understanding which factors drive health insurance costs is beneficial for individuals who are selecting or planning changes to their coverage, moreso when considering having children or relocating. With knowledge, consumers can make informed financial decisions and manage future expenses.

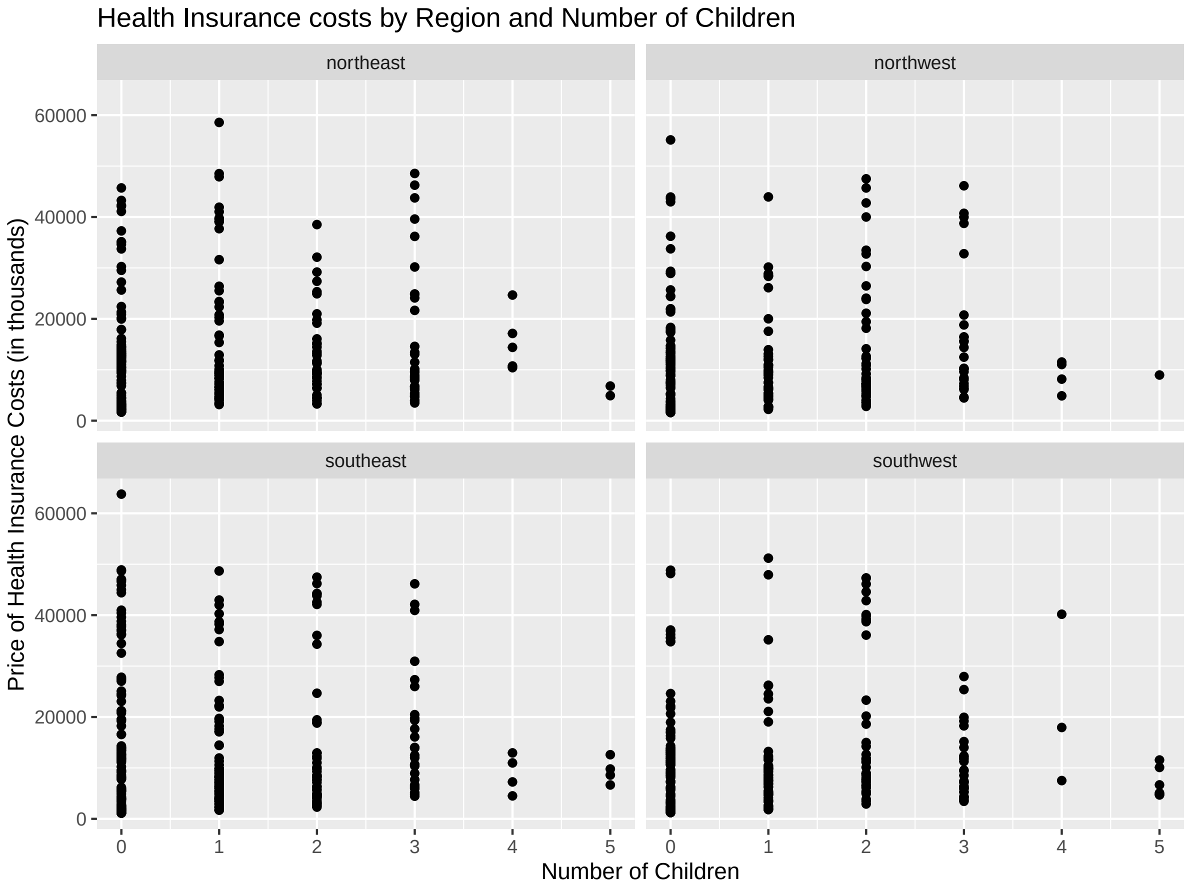
Based on background knowledge, this leads to some hypothesis: 1) Region and price of health insurance does not have any relationship 2) The number of children has no correlation with price of health insurance 3) The effect of the number of children individuals have on the price of health insurance does not depend on the region they live in.

Data



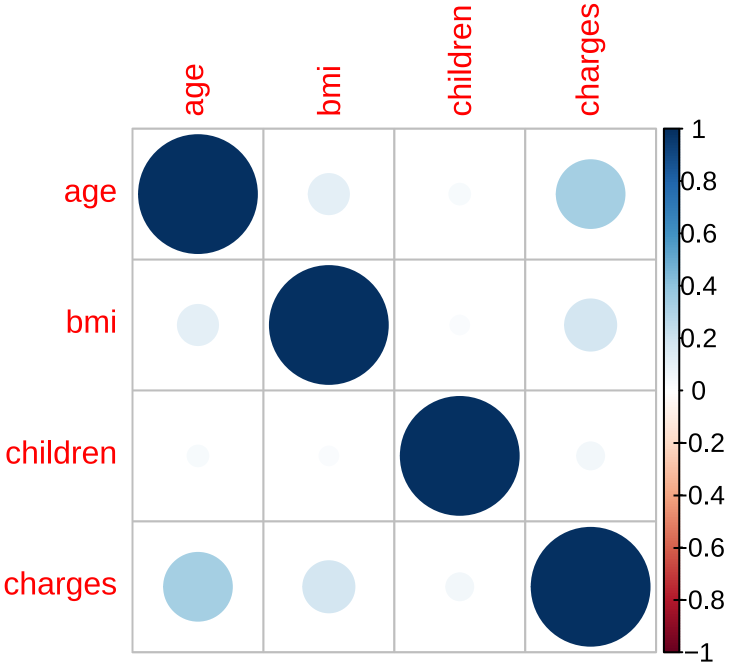
This graph looks at region as an explanatory variable to see whether location impacts the price of health insurance. From the boxplot, we can see that all four regions show very similar outcomes, with most median insurance costs falling roughly between 8,000 and 10,000. This suggests that location alone does not drastically change typical insurance prices. For someone considering moving to a new region, this is useful because it shows that insurance cost is not heavily tied to where you live.

However, the plot also shows a noticeable spread within each region, meaning there are other important factors influencing insurance cost beyond location. The minimum values across regions fall around 1,000–2,000, while the 25th percentile is near 5,000. The 75th percentile varies more, with the southwest appearing slightly lower and the southeast slightly higher. The same pattern holds for maximum values. Overall, while a slight difference exists such as the southwest trending lower and the southeast trending higher, the similarities across regions suggest that location is not the main influence of health insurance cost.



This graph examines how both region and number of children relate to health insurance cost. Each panel represents one region, allowing us to compare patterns across locations. Overall, we can see that as the number of children increases, insurance costs tend to remain similar with some decreases in every region, although the amount of variation within is a bit spread. The southeast appears to show slightly higher insurance costs on average, while the southwest generally appears somewhat lower, but the differences between regions are not dramatic. For families with four or more children, the sample size becomes much smaller, so any conclusions for those groups should be interpreted cautiously. Overall, this plot suggests that the number of children has a more noticeable effect on insurance cost than region does, but both variables contribute in some way.

```
children <- insurance |> select_if(is.numeric)

corrplot(cor(children))
```



Based on the correlation plot, the correlations between charges and number of children appear pretty weak, indicating that children count does not show a strong linear association with health insurance costs. Most values on this plot are close to zero, supporting the idea that insurance charges may depend on other factors not included in this study.

Table 1: Summary of Health Insurance Costs by Region and Number of Children

region	children	mean_cost	median_cost	sd_cost	n
northeast	0	12108.561	10435.065	10920.129	107
northeast	1	16654.844	9617.662	14252.471	57
northeast	2	13074.272	10043.249	8590.251	39
northeast	3	14804.673	8606.217	13131.047	37
northeast	4	15467.689	14394.398	5844.089	5
northeast	5	5857.259	5857.259	1332.471	2
northwest	0	11052.764	8835.265	10652.659	103
northwest	1	10757.748	7518.025	9346.859	49
northwest	2	13962.802	8116.269	12112.069	49
northwest	3	16848.501	12475.351	12577.909	25
northwest	4	8898.615	9598.189	3062.551	4
northwest	5	8965.796	8965.796	NA	1
southeast	0	14249.204	9212.256	14699.685	124
southeast	1	13207.233	8233.097	11917.529	71
southeast	2	14714.643	8219.204	14498.099	47
southeast	3	16467.400	13202.812	11718.824	28

region	children	mean_cost	median_cost	sd_cost	n
southeast	4	8918.709	9110.510	3775.844	4
southeast	5	9408.004	9192.847	2485.290	4
southwest	0	11241.816	9630.229	10445.736	102
southwest	1	10207.341	7624.630	9993.433	61
southwest	2	17780.402	11150.780	15014.031	43
southwest	3	10191.662	7418.522	6587.394	29
southwest	4	21878.873	17942.106	16686.985	3
southwest	5	7152.440	5873.169	2967.511	6

The table above summarizes health insurance costs by both region and number of children, displaying mean, median, standard deviation, and sample size for each group. Overall, the average insurance cost fluctuates up and down as the number of children increases across all regions, supporting the trend observed in the scatterplot. For example, in both the northeast and southeast regions, the mean cost fluctuates from families with zero children to families with three children, and then decreases drastically when looking at families with four or five children. While there are small regional differences, they are not dramatic; the southeast consistently shows slightly higher average costs, while the southwest tends to show slightly lower averages. The standard deviation values across groups are quite large, showing that insurance costs vary widely within each region and child count category. This high variability suggests that, beyond location and number of children, other factors such as age, BMI, and smoking status likely play a major role in determining insurance cost. In general, the table reinforces the idea that the number of children has a stronger and more consistent association with insurance costs than region alone, while still showing variation across each region.

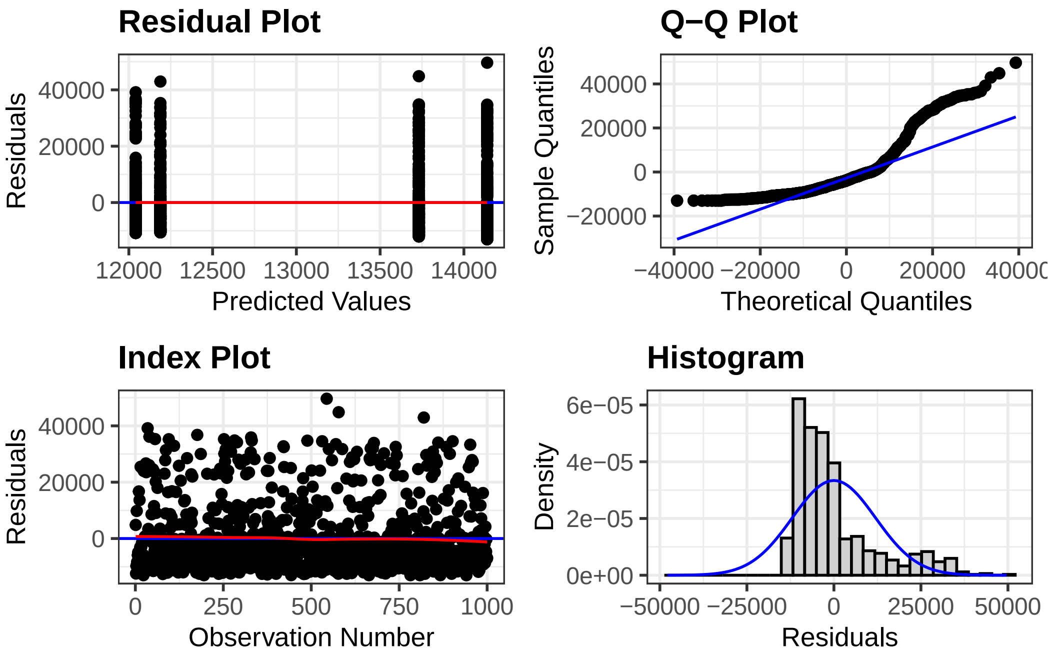
Models

For the 1st model I chose region as the explanatory variable because it's known that health insurance costs tends to vary from place to place, and it seemed reasonable to explore whether region alone contributes to difference in charges. Also the way financial systems and pricing works is different across locations so assuming same cost across all regions would not be appropriate. The mathematical equation for the model is:

$$\widehat{\text{charges}} = \beta_0 + \beta_1 \times \text{regionNW} + \beta_2 \times \text{regionSE} + \beta_3 \times \text{regionSW}.$$

I fitted the model in R using the lm() function, then reviewed the summary output. Upon examining the assumptions, the residual plot showed no curve, suggesting linearity is acceptable, but the vertical grouping shows potential violation in constant variance. The QQ plot showed a great amount of divergence from the diagonal line on both ends, violating normality assumption. The index plot showed random scatter, satisfying the independence, and the histogram showed right-skewedness, meaning the model tends to underpredict high charges. I didn't include any transformations, nonlinear terms, or interactions in this model because I wanted to see the effect of region alone towards insurance costs.

```
Call:
lm(formula = charges ~ region, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max 
-13017  -8455  -3860   2931  49632 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      13731.4      761.5  18.032   <2e-16 ***
regionnorthwest  -1543.1     1095.4  -1.409    0.159    
regionsoutheast    407.4     1046.5   0.389    0.697    
regionsouthwest  -1690.5     1080.2  -1.565    0.118    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11970 on 996 degrees of freedom
Multiple R-squared:  0.006017,	Adjusted R-squared:  0.003023 
F-statistic:  2.01 on 3 and 996 DF,  p-value: 0.1109
```

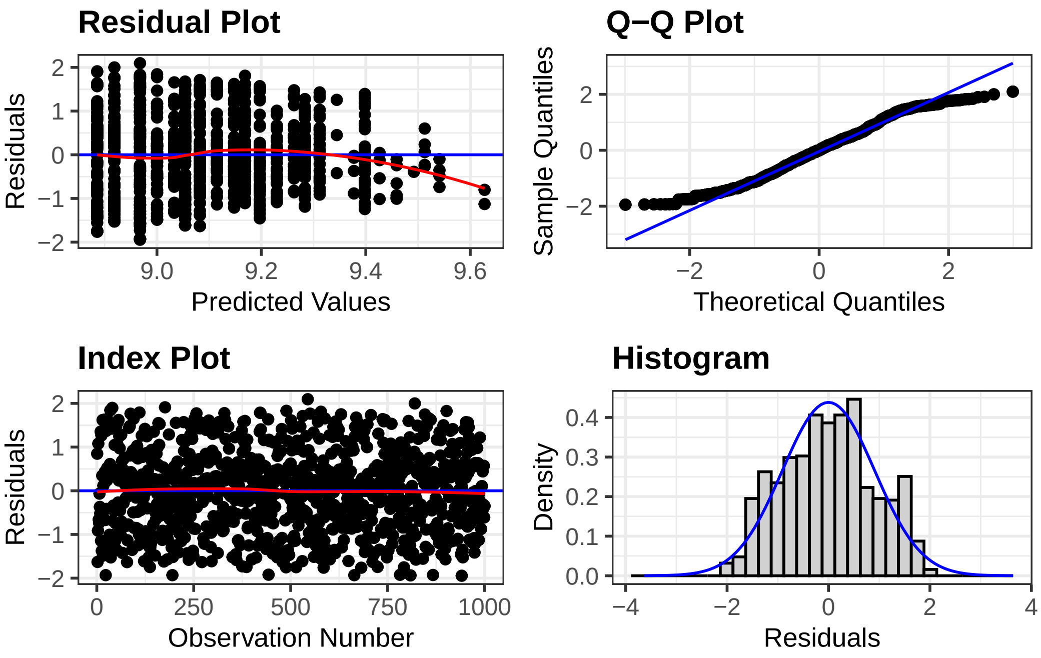


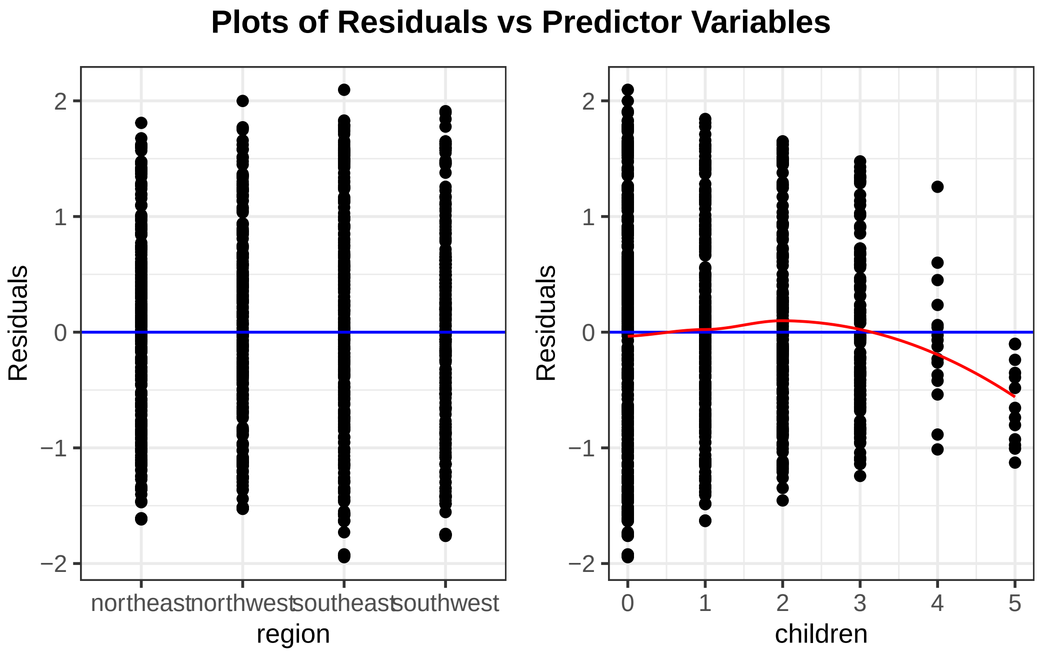
For the 2nd model, I added number of children as an explanatory variable to see if this factor could better explain the variation in insurance costs. Additionally, since Log transformation is known to help with normality violations (right-skewed data), I proceeded to apply it to model 2 so it can address that concern. The equation for the model is:

$$\log(\text{charges}) = \beta_0 + \beta_1 \times \text{regionNW} + \beta_2 \times \text{children} + \beta_3 \times \text{regionSE} + \beta_4 \times \text{regionSW}$$

I fitted this model in R, where the log transformation improved normality greatly as seen in the symmetry/bell-shape from the histogram and the closer points to the diagonal line on the QQ plot. However, there is still a slight funnel shape (heteroscedasticity), and the residuals for the children plot showed a clear non-linear trend at the far right, giving us the assumption of constant variance and linearity still being a concern.

```
Call:
lm(formula = log(charges) ~ region + children, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.94484 -0.68989 -0.00096  0.60353  2.09545 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      9.05398    0.06397 141.524  < 2e-16 ***
regionnorthwest -0.13546    0.08350  -1.622   0.1050    
regionsoutheast -0.08639    0.07979  -1.083   0.2792    
regionsouthwest -0.16841    0.08233  -2.046   0.0411 *  
children         0.11470    0.02409   4.761 2.21e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9121 on 995 degrees of freedom
Multiple R-squared:  0.02686,	Adjusted R-squared:  0.02295 
F-statistic: 6.866 on 4 and 995 DF,  p-value: 1.879e-05
```





```
  (Intercept) regionnorthwest regionsoutheast regionsouthwest        children
 8552.5263292       0.8733100       0.9172374       0.8450106       1.1215404
```

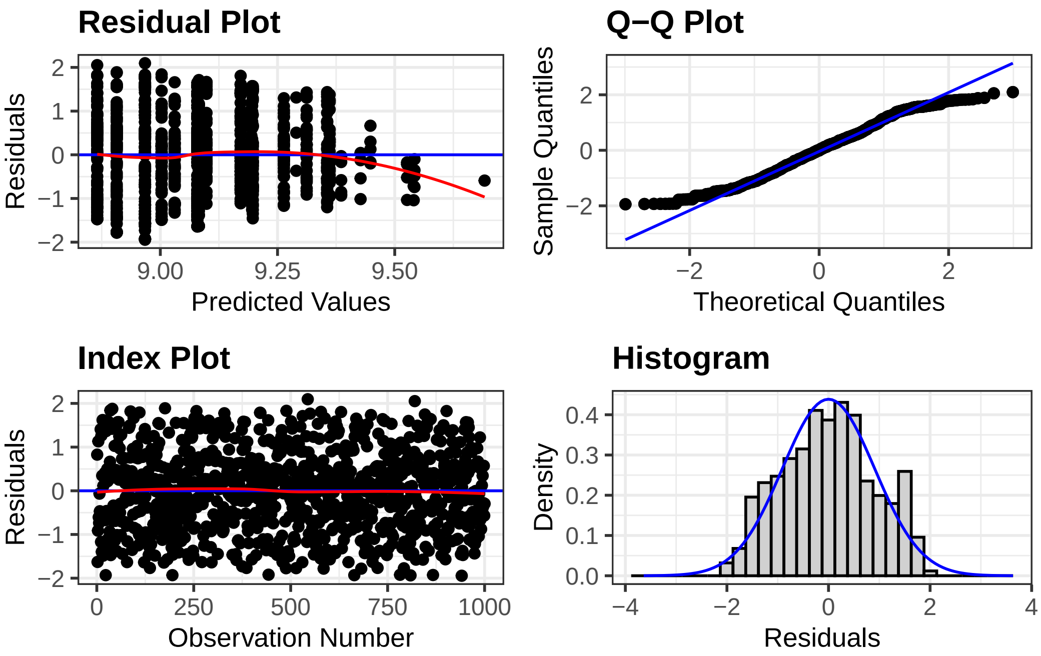
```
            2.5 %   97.5 %
children 1.069756 1.175832
```

For the 3rd model I kept region and number of children, log-transformation, and added interaction between region and number of children to see if the effect of having more children on insurance costs depended on the specific region. The equation for model 3 is:

$$\log(\text{charges}) = \beta_0+\beta_1\times\text{regionNW}+\beta_2\times\text{regionSE}+\beta_3\times\text{regionSW}++\beta_4\times\text{children}+\beta_5\times(\text{regionNW}\times\text{children})+\beta_6$$

I again fitted the model in R, where the assumption checks for model 3 were nearly identical to those of model 2, while also still showing good normality but some concerns about constant variance and linearity.

```
Call:
lm(formula = log(charges) ~ region * children, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.94458 -0.69505  0.00149  0.60873  2.09571 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               9.079031   0.078769 115.261   <2e-16 ***
regionnorthwest          -0.213710   0.113366  -1.885   0.0597 .  
regionsoutheast          -0.111692   0.107192  -1.042   0.2977    
regionsouthwest          -0.172023   0.111478  -1.543   0.1231    
children                  0.092286   0.047614   1.938   0.0529 .  
regionnorthwest:children  0.072986   0.070810   1.031   0.3029    
regionsoutheast:children  0.022665   0.066663   0.340   0.7339    
regionsouthwest:children  0.003469   0.066792   0.052   0.9586    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9129 on 992 degrees of freedom
Multiple R-squared:  0.02815,	Adjusted R-squared:  0.02129 
F-statistic: 4.105 on 7 and 992 DF,  p-value: 0.0001862
```



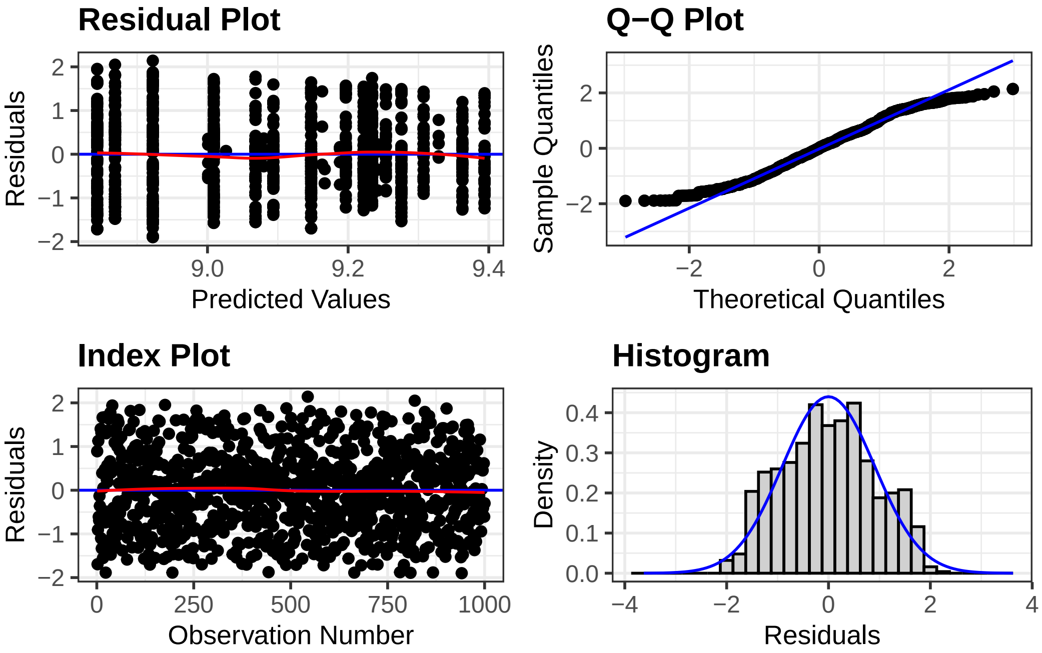
For the 4th model I decided to keep region and children, take out the interaction between the two, and instead add a quadratic term for children to resolve the linearity assumption violation. The equation for model 4 is:

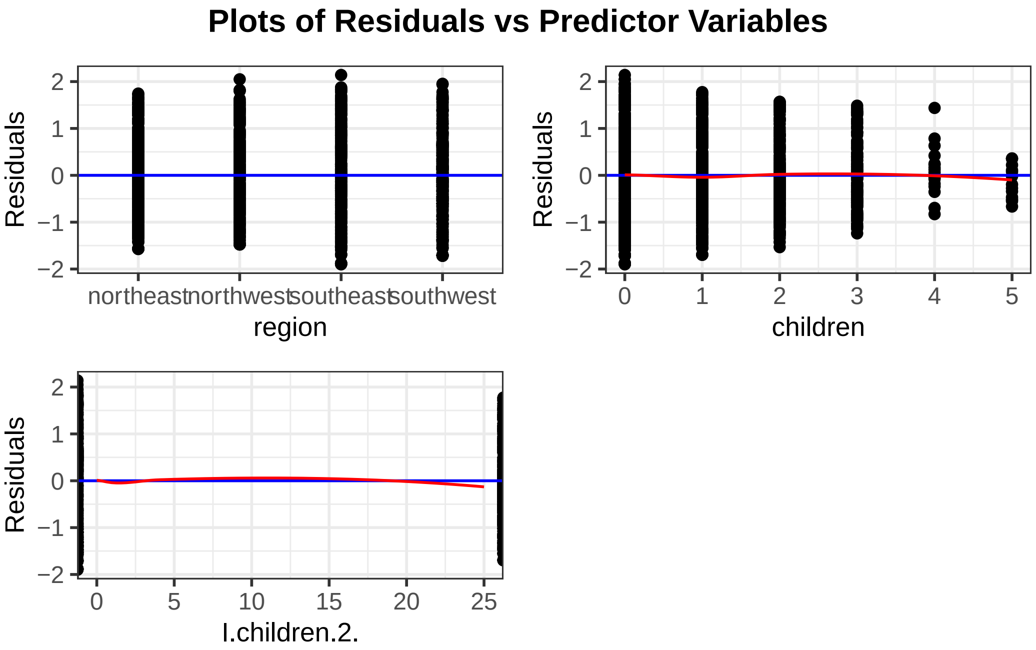
$$\log(\text{charges}) = \beta_0 + \beta_1 \times \text{regionNW} + \beta_2 \times \text{regionSE} + \beta_3 \times \text{regionSW} + \beta_4 \times \text{children} + \beta_5 \times \text{children}^2$$

I fitted this model in R, where now if we look at the assumptions, each remain the same, besides the residual plot; it now doesn't violate the linearity assumption, but still violates the constant variance assumption.

```
Call:
lm(formula = log(charges) ~ region + children + I(children^2),
    data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.89947 -0.67584  0.00369  0.63002  2.14082 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      9.00891    0.06581 136.883  < 2e-16 ***
regionnorthwest -0.14040    0.08324  -1.687  0.09197 .  
```

```
regionsoutheast -0.08668    0.07952  -1.090  0.27596
regionsouthwest -0.16579    0.08206  -2.020  0.04362 *
children         0.27361    0.06230   4.392 1.24e-05 ***
I(children^2)   -0.04841    0.01751  -2.764  0.00581 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9091 on 994 degrees of freedom
Multiple R-squared:  0.03429,	Adjusted R-squared:  0.02943
F-statistic: 7.058 on 5 and 994 DF,  p-value: 1.703e-06
```





```
   (Intercept) regionnorthwest regionsoutheast regionsouthwest        children 
  8175.5889084       0.8690063       0.9169693       0.8472276       1.3147014 
 I(children^2) 
     0.9527434 
```

```
                   2.5 %    97.5 %
I(children^2) 0.9205571 0.9860551
```

Results

Firstly, it was rather surprising finding out that Region as an explanatory variable barely explains the variability in price of health insurance. Looking at model 1, based on the summary of R, we obtained an R^2 of 0.006, which is roughly 0.6% of the variability in price of health insurance. Not only that but we can see that each region has a p-value over 0.05, meaning that there's no strong evidence that region alone has an effect on insurance price. Additionally our assumptions showed constant variance and normality violations so adding log transformation was part of the next step.

In model 2, it was interesting that applying log transformation and number of children as an explanatory variable in addition to region, that our $R^2 = 0.026$, which is 2.6%. The number of children became significant (p-value approximately 0), telling us that families with more children tend to have higher insurance costs (21% increase). While most region's had a p-value over 0.05, southwest this time had a p-value of 0.04, showing significantly lower average insurance costs than the northeast region, telling us that taking number of children into account helped explain additional variability, a more accurate estimate. Based on the 95% confidence interval, we see that for each additional child, health insurance cost is estimated to increase by approximately 7% to 18%, as the multiplicative effect is 1.07 to 1.18. Adding log transformation improved normality, but some heteroscedasticity remained, probably because of the uneven spread across predicted values and the small number of observations for families with four or five children.

In model 3, after adding an interaction term we see that the p-value of northwest region alone lowered to 0.05, all others increased above 0.05, and as for children, the p-value increased, but remained at 0.05. Additionally, each interaction between region and children had a p-value above 0.05, and our R^2 value increased very slightly to 0.028, a 2.8% compared to model 2's 2.6%. This tells us that the interaction term is not significant, providing no strong evidence that the effect of children on insurance costs differs by region. It only added complexity, therefore model 2 is a better fit.

Lastly, in mode 4, after adding a quadratic term to children, we see that linearity is addressed and no longer violated. Not only that, but if we look at the p-value for children^2, it is less than 0.05, meaning the quadratic term is very significant. Furthermore, if we look at our R^2 value, it now increased to 0.03429, which is 3.4%, which is an even bigger increase compared to model 2's R^2 value of 0.026 (2.6%). Diving deeper, the positive linear term for children and negative quadratic term for children^2 indicates a non-constant, non-linear relationship. This tells us that insurance increases with the first few children, but the rate of increase slows down as the number of children gets higher. Also, the southwest region remains significant (p = 0.04) and is associated with lower health insurance costs compared to the northeast baseline, while the northwest and southeast regions show no significant difference. Based on the 95% confidence interval, we are 95% confident that for every one-unit increase in children^2, the average health insurance cost is multiplied by a factor between 0.921 (7.9% decrease in the rate of change) and 0.986 (1.4% decrease in the rate of change). Overall, because model 4 addresses the non-linearity assumption and gives a higher R^2, it is deemed the better model out of the four for this analysis.

Conclusion

Overall, it seems that Region and Number of Children barely explained any variability in price of health insurance. Although the number of children showed a statistically significant positive relationship with insurance cost, the best-fitting Model 4 still had a weak overall fit, showing only 3.4% of variability. Furthermore, we found no meaningful interaction between region and children, supporting the hypothesis that the effect of family size does not depend on location. Crucially, Model 4 revealed that the relationship between children and cost is non-linear: costs increase with the first few children, but the rate of increase diminishes as the number of children rises.

For medical professionals and policymakers, this study highlights that family size and regional differences tend to be poor predictors of health insurance costs. The low R^2 values tell us that there are other unexamined factors such as age, sex, BMI, and smoking status that potentially explain more variability in determining health insurance prices. All things considered, our conclusions are limited by the small sample sizes for families with four or five children and the unresolved concern regarding the constant variance assumption in the models. Specifically, the constant variance violation in the residual plot means the standard errors and p-values reported may be less reliable. To create more precise or accurate models/fairer policies, stakeholders should consider other additional factors as those are more likely to explain a larger portion of cost variation.

Citations/References

ChatGPT (word structuring/clarification).

https://www.kaggle.com/datasets/mirichoi0218/insurance